Design of Sentiment Analysis System Based on Multilingual Social Network Data

SHORT SYNOPSIS
Submitted in partial fulfillment of the requirement of the degree of

DOCTOR OF PHILOSOPHY

to

Manav Rachna International Institute of Research and Studies
(Deemed to be University)

by

Neha Garg
Registration Number: 17/Ph.D/051

Under the supervision of

Dr. Kamlesh Sharma
Associate Professor, CSE
Faculty of Engineering & Technology
Manav Rachna International Institute of Research and Studies

FACULTY OF ENGINEERING & TECHNOLOGY

MANAV RACHNA INTERNATIONAL INSTITUTE OF RESEARCH & STUDIES

SECTOR-43, SURAJKUND –DELHI ROAD, FARIDABAD – 121001

FACULTY OF ENGINEERING & TECHNOLOGY

Certificate

I hereby declare that the short synopsis "Design of Sentiment Analysis System Based on Multilingual Social Network Data" being submitted in partial fulfillment of the requirement for the degree of Doctor of Philosophy in Computer Science and Engineering under Faculty of Engineering & Technology of Manav Rachna International Institute of Research & Studies, Faridabad during the academic year 2019.

It is a bonafide record of original work carried out under the supervision of Dr. Kamlesh Sharma, Associate Professor, CSE, FET and has not been submitted in full or part to any other university or institute for degree or diploma. Further, it is certified that short synopsis is plagiarism free and the proof has been submitted.

(Neha Garg)
17/Ph.D/051

Signature of the Supervisor
Dr. Kamlesh Sharma
Associate Professor,CSE, FET
Manav Rachna International Institute of Research and Studies

(Chairman, DRC)

Chairperson, CDP

ABSTRACT

The usage of social sites is increasing among a huge variety of users like activist, spammer, toll, etc and the advancement of the World Wide Web (WWW) is enabling users to post data in their structure (video, audio, text etc.) and language (English, Hindi, Spanish, French, German etc.). With such flexibility, more users are getting associated with social networking sites and posting their opinions on the latest issues for example performance of BJP in the center, etc. And this gives birth to a new trend of multilingual communication among the communities. People are expressing themselves in the languages; they are comfortable with and mixing the words of different languages by using phrases etc and switching the language in the single sentence to make their blogs, tweets, speech, etc. more impactful. The situation where more than one language is used in the context of a single issue in a single utterance is called Linguistic Code-Switching, on the other hand when users use morphemes or phrases of one language in the other language, it is called Code Mixing. Since customers are expressing their views by twitting or posting on every event or raise issues related to their surroundings and environment, to represent their opinions, these opinions can provide feedback or information related to that field which can be used further for sentiment analysis.

Researchers are working on sentiment analysis of the multilingual communication including Hindi, English, Telugu, etc.; but the switching between Hindi and English i.e. Hinglish is not taking into account because of various constraints such as ambiguity, inconsistent sentence structure, inconsistent spellings, using abbreviations, emoticons, etc. imposed and considering it as a noise. While a proper translation of Hinglish can make the analysis more impactful.

Further, at a particular time, many bilingual and cross-lingual events may occur in parallel related to a single issue that can affect the targeted event. The analysis of these events can be done by making clusters of events and make nodes based on popularity and similarities. In this scenario, nodes having similarities are having more chances of being connected irrespective of their popularity. By analyzing all these events and their sub-events, the more accurate results will be found which will result in strengthening the decision making.

In this research, after considering all these challenges and issues, the proposed model will consider Hindi; English and Hinglish words involved in the communication create clusters of

similar events and the related sub-events, perform sentiment analysis and use the result of the sentiment analysis for better prediction.

Keywords: Big Data, Big Data Analytics, Machine learning, Code Mixing, Linguist Code Switching, Predictive Analysis, Sentiment Analysis, Multilingual Communication, Deep Learning, Hinglish.

CONTENTS

S. No.	Description	Page No.
1	Introduction	1-9
	1.1 Predictive Analysis	
	1.2 Sentiment Analysis	
	1.3 Code- Mixing	
	1.4 Linguistic Code-Switching	
	1.5 Deep Learning Algorithm	
2	Literature Review	9-12
3	Description of Broad Area	12-14
	3.1 Research Challenges	
	3.2 Problem Formulation	
4	Research Objectives	14
5	Research Methodology	14-16
6	Expected Outcome of the Research	16
	References	17-20

1. INTRODUCTION

It's a competitive world. The competition is a key to innovation and discovery. Due to the competition, the customer is getting the desired item at a reasonable price and good quality. To sustain and to progress, it is inevitable for any organization to provide the product to the market before the competitors do. Customer feedback plays a very important role in this which enables one to establish demand. With decision making an analysis based on customer feedback, prediction of future demand of the market and new product can be introduced accordingly. The accuracy of decision making analysis is what makes any organization be at a pole position or doom (Tripathy, A., Agrawal, A., and Rath, S.K., 2015). As time has progressed, customer feedback is not the only source of customers' opinions. The use of social media (SM) such as comments, likes, using emoticon etc on Facebook, tweets on Twitter and audio and video on Instagram etc. by an individual is another mode where personal choices and interests are expressed (Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C. and Byers, A.H., 2011) in all forms of data (Sagiroglu, S. and Sinanc, D., 2013). Similarly, for an event, the users of social media are posting blogs, tweets, etc. to express their interest and feelings. This social media data can be extracted and put to use for decision making. With the increasing usage of social media (SM) by various kinds of users in an unpredictable manner, data is generated at a very rapid pace (Gandomi, A. and Haider, M., 2015; Beyer, M.A., and Laney, D., 2012). By using this huge amount of information, available on social sites, the system enhances the worth of decision making by improving the outcomes, extracted from databases (Lesser, V., Horling, B., Klassner, F., Raja, A., Wagner, T. and Zhang, S.X., 2000).

Hence the various methods to handle all forms of data have been introduced, which are discussed in the following section.

1.1 Predictive Analysis

The predictive analysis comprises various techniques to forecast future outcomes based on historical data. The data can be of any form structured (age, name, etc), unstructured (posts on social media, etc) or semi-structured like social sites, which need to be extracted in meaningful information. The predictive analysis seeks to uncover patterns and relationships

among these data and provide information on which one can take the decision. The predictive analysis can be classified in two ways-

A. Regression Techniques

B. Machine Learning Techniques

A. Regression Techniques

This technique mainly focuses on developing a mathematical equation for modeling the relationships among the various variables under the concern. Using Regression and the problem in hand, a vast range of models can be developed to perform predictive analytics. A few models are discussed in Table 1.

Table 1: Types of Regression Techniques

Techniques	Algorithm Type	Description
Parametric Techniques	Linear Regression Technique (Zou, K.H., Tuncali, K. and Silverman, S.G., 2003)	• This technique develops equation between the resultant variable (dependent variable) and a set of independent (predictor variables). • This equation predicts the resultant variable as a linear function of the parameters.
	Discrete Choice Models (Greene, W., 2009)	• As the name suggests when the dependent variable has discrete values. • Some examples of these methods are- - logistic regression, (generally used when dependent variables are binary) - multinomial logit and - probit models(generally used when dependent variables are binary).

	Time Series Models (Bhaskaran, K., Gasparrini, A., Hajat, S., Smeeth, L. and Armstrong, B., 2013)	• These types of models generally take into account the data points which have some internal correlation over time like the weather forecast. • The dynamic path modeling of a variable is used to enhance the quality of forecasting results because the predictable components of the series can be estimated in the future. Examples are- - Moving-Average (MA) models and - Autoregressive models (AR).
Non-Parametric Techniques	Classification And Regression Trees (CART) (Lewis, R.J., 2000)	• A Decision tree technique for predicting consistent dependent variables (regression) and categorical predictor variable (classification) and forms a tree based on certain rules.
	Multivariate Adaptive Regression Splines (MARS) (Friedman, J.H. and Roosen, C.B., 1995)	• It develops flexible models by fitting piecewise linear regressions. • MARS is based on the concept of Knot, in which the local regression model provides the path to other models, that's how to provide the intersection point between two splines.

B. Machine Learning Techniques

The main aim of the machine learning techniques is to make the machines reply, without depending upon the program develops in advance for that particular problem (Mujeeb, S. and Naidu, L.K., 2015). Machine Learning Techniques are defined as follows.

i. Supervised Algorithm

Supervised algorithms mean that a program is made or modeled on a predetermined set of sample data, which helps in finding an approximate output when new data is supplied in place of the predetermined dataset (Bringmann, B., Berlingerio, M., Bonchi, F. and Gionis, A., 2010). Various types of supervised algorithms are tabulated in Table 2.

Table 2: Categories of Supervised Algorithm

S. No.	Algorithm Type	Description	Input values	Example
1.	Regression Algorithms (Doan, T. and Kalita, J., 2015)	• These algorithms formed a model to estimate the relationship between dependent and independent variables. • With the help of multiple such examples, a model/function is developed which helps in decision making.	dependent and independent variables	• Ordinary Least Squares Regression (OLSR), • Linear Regression, • Logistic Regression, • Stepwise Regression
2.	Instance-based Algorithms (Beringer, J. and Hüllermeier, E., 2007)	• In this algorithm already available example-data is examined to model the relationship with the current targeted value. • So, it's using a measure to find similarities, to get optimal matches and to make the prediction. • For example, speller correction, search	Predefined data set	• k-Nearest Neighbor (kNN), • Learning Vector Quantization (LVQ), • Self-Organizing Map (SOM)

		ranking, etc. can be solved under this algorithm.		
3.	Regularization Algorithms (Bauer, F., Pereverzev, S., and Rosasco, L., 2007)	• If the prediction performance is not so good then the issue is called the overfitting problem. This can be removed by Regularization algorithms.	The finite set of variables	• Ridge Regression, • Least Absolute Shrinkage and Selection Operator (LASSO), • Elastic Net, Least-Angle Regression (LARS).
4.	Decision Tree Algorithms (Cord, M. and Cunningham, P. eds., 2008)	• A decision tree is a binary form for a set of attributes to be tested to predict the output. • The nodes of the tree split the data to find a classifying variable. • This algorithm handles the problem logically and does step-wise execution to get good results.	Set of attributes either having continuous values or binary values.	• Classification and Regression Tree (CART), • Iterative Dichotomies 3 (ID3), • C4.5 and C5.0 (different versions of a powerful approach), • Conditional Decisi on Trees.
5.	Bayesian Algorithms (Domingos, P. and Pazzani, M.,	• Those algorithms that incorporate Bayes Theorem. • The Bayesian method	Expected probability	• Naive Bayes, • Gaussian Naive Bayes, • Multinomial

	1997)	means the one which believes in subjective probability and results in some future reference.		Naive Bayes.
6.	Artificial Neural Network Algorithms(Cord , M. et al. 2008)	• The neural network aims to solve the problem in the same way as solved by the human brain with intelligence. • Hence it is used in ML. • The area is much related to pattern matching problems.	Single neurons with the arbitrary number of inputs.	• Preceptor, • Back-Propagation, • Hopfield Network.
7.	Deep Learning Algorithms(Cord , M. et al. 2008)	• Deep learning algorithms are among the special cases for Artificial Neural Networks. • They help to build more complex neural networks and work for voluminous datasets.	Use a cascade of multilayer of non-linear processing units. Each layer takes the output of previous layers as input.	• Deep Boltzmann Machine (DBM), • Deep Belief Networks (DBN), • Convolution Neural Network (CNN)

ii. Unsupervised Learning Algorithm

Unsupervised algorithms mean that a program is provided with some collection of data, with no predetermined dataset being available, and one has to make patterns or relations among those data values. The data given to unsupervised algorithms are not labeled, which means

the input variables are given with no specified output variables. In this type of algorithms, one has to find the pattern according to problem definition. The following are different types of unsupervised algorithms in Table 3.

Table 3: Categories of Unsupervised Algorithms

S.No.	Algorithm	Description	Example
1.	Clustering Algorithms (Musumeci, F., Rottondi, C., Nag, A., Macaluso, I., Zibar, D., Ruffini, M. and Tornatore, M., 2018)	Clustering algorithms help in creating a cluster/group of similar elements/data/problems	• k-Means, • k-Medians, • Expectation-Maximization (EM), • Hierarchical Clustering.
2.	Dimensionality Reduction Algorithms (Cortés, G., Benitez, M.C., García, L., Álvarez, I. and Ibanez, J.M., 2015)	In dimensionality reduction, the number of dimensions of the internal structure of data is reduced without harming the actual data content.	• Principal Component Analysis (PCA), • Principal Component Regression (PCR), • Partial Least Squares Regression (PLSR)
3.	Association Rule Algorithms (Musumeci, F. et al., 2018)	finds the relationship between dataset and data	• Apriori algorithm, • Éclat algorithm.

1.2 Sentimental Analysis

The process of identifying and categorizing the opinion of the user towards an entity expressed in the form of text; this is especially used to determine the attitude of the user. With the rapid increase in using social media and users are expressing themselves with the help of textual data in form of blogs, reviews, rating etc. there is a sudden need to analyze

their text based on sentiment i.e. if the content is in favor of our business or not (Tripathy, A., Agrawal, A., and Rath, S.K., 2015).

For predicting the behavior of humans against one's organization, machine learning algorithms are very useful. The supervised algorithm utilizes a named set of training data which is record against a suitable sentiment. Though, unsupervised learning incorporates dataset where content is not recorded with proper sentiments (Singh, Y., Bhatia, P.K. and Sangwan, O., 2007; Tripathy, A., Agrawal, A., and Rath, S.K., 2015; Bringmann, B. et al., 2010).

Sentimental analysis can be done in following manners-

a) ***Document Level***- where the whole document is considered and then the classification of sentiments either it is positive, negative or neutral is made.

b) ***Sentence Level-*** in this category polarity of sentiments is decided based on the individual sentences. Hence there is a possibility of multiple polarity representation towards an entity in a single document.

c) ***Aspect Level-*** In this category classification of sentiments is done for the particular aspects of an entity.

1.3 Code –Mixing Phenomenon

Code-Mixing (CM) can be defined as the implanting of semantic units, for example, expressions, words, and morphemes of one language into an expression of another language, which is commonly experienced in multilingual communities (Myers-Scotton, C. 1997; Poplack, S., 1980; Muysken, P., Díaz, C.P. and Muysken, P.C., 2000; Chandu, K., Loginova, E., Gupta, V., van Genabith, J., Neuman, G., Chinnakotla, M., Nyberg, E. and Black, A.W., 2018). This sometimes makes a feels of lack of fluency in the segment of either language. However, with the WWW support to multiple languages, code-mixing has become such a general task among the multi-lingual community that everyone uses multiple languages to powerfully expressing their views, in their texts intentionally and unintentionally.

1.4 Linguist Code –Switching

To express one's views impressively and to change phonetics while communicating to produce more impactful speech, people commonly switch from one language to the other

language in a single utterance. In India, a vast community switches to English while talking into Hindi and vice versa (Huang, F. and Yates, A., 2014). This represents a situation where two and more languages are being used in a single conversation. E.G. in Hindi and English code ***" tumne Newspaper padh liya ya ni", "Mei movie dekhne jar hi hu by car"***. Traditionally machine learning models treat LCS as noise but by carefully analyzing the text one can find more fruitful information from it.

1.5 Deep Learning Algorithm

Deep learning algorithms are the special cases of Artificial Neural Networks. They help to build more complex neural networks and work for voluminous datasets. Use a cascade of multilayer of non-linear processing units. Each layer takes the output of previous layers as input. (Bringmann, B. et al., 2010) Here in this research, the proposed model will develop clusters of various events and make nodes connected to these clusters by checking similarity and popularity parameters, using deep learning parameters (Papadopoulos, F., Kitsak, M., Serrano, M.Á., Boguná, M. and Krioukov, D., 2012).

2. LITERATURE REVIEW

The stream of data is being generated from various users and is of various forms, in such a heterogeneous environment, data management is become very difficult using traditional data management systems (Phaneendra, S.V., and Reddy, E.M., 2013), analysis system and warehouse. After the Hace theorem, (Tamhane, D.S., and Sayyad, S.N., 2015) applied the concept of big data on heterogeneous data and on the heterogeneous users and the era switched from a centralized structure to decentralized structures. The dimensions of big Data i.e. V's are introduced by many organizations e.g. IBM, Oracle, SAS, etc from time to time; velocity, variety, and volume are introduced (Doug., L. 2001; Gandomi, A. and Haider, M., 2015; Beyer, M.A., and Laney, D., 2012), viscosity and virality (Mujeeb, S. and Naidu, L.K., 2015.), Veracity By IBM, value by Oracle and ***variability*** by SAS (Gandomi, A. and Haider, M., 2015; Fan, J., Han, F. and Liu, H., 2014.; Diebold, F.X., Cheng, X., Diebold, S., Foster, D., Halperin, M., Lohr, S., Mashey, J., Nickolas, T., Pai, M., Pospiech, M. and Schorfheide, F., 2012.). Later on, organizations are introducing V's according to the need of their organizational structure. Currently, the world of big data has been populated with the idea of 45 V's, which are flexible enough and can be adopted by organizations with ease.

As the heterogeneity in data types introduced, various techniques for analysis of different kinds of data are coming into play. Text Analytics (Jha, A., Dave, M. and Madan, S., 2016; Gandomi, A. and Haider, M., 2015) introduced to change a huge amount of text data into useful information for decision making using statistical analysis, computational semantics and machine learning techniques. Audio Analytics (Gandomi, A. and Haider, M., 2015) is helpful in speech analysis of unstructured data. Video Analytics (Elgendy, N. and Elragal, A., 2014) is helping in the area of security and surveillance.

Social Media Analytics (Elgendy, N. and Elragal, A., 2014) is the current trend where one needs to handle all kinds of data and Predictive analytics is creating a new era for forecasting based on the historical data.

However, according to the business needs and to make accurate decision the organizations are doing predictive analyses, researchers using different techniques to solve their problems according to their problem definitions for example by making a mathematical equation based on relationships among various variables called regression models, using historical data and some predefined parameters which may results in supervised algorithms or without using any predefined parameters, called unsupervised learning (Bringmann, B., Berlingerio, M., Bonchi, F. and Gionis, A., 2010).

Pang, B., Lee, L. and Vaithyanathan, S., (2002) have done sentiment classification based on the experiments using machine learning algorithms i.e., Naive Bayes classification, Support Vector Machine, and Maximum Entropy classification and categorized it as positive and negative sentiments.

Turney, P.D., (2002) considered, unsupervised algorithm, Using Part of Speech (POS) technique to find phrases, contain adjectives or adverbs to identify movie review in two groups- prescribed i.e., Thumbs up and not suggested i.e., Thumbs down.

Dave, K., Lawrence, S. and Pennock, D.M., (2003) utilized structured review for testing and training, recognizing features, score strategies, and inverse document frequency to weigh the occurrences of words and based on it; determine reviews as positive or negative.

Pang, B. and Lee, L., (2004) have labeled sentences in the document by considering the subjective group only which results in cleaner and lesser data for polarity classification using

Naïve Bayes and Support Vector machine techniques. This technique represents that applying the minimum-cut framework can result in more efficient algorithms.

Whitelaw, C., Garg, N. and Argamon, S., (2005) have given a sentiment classification technique based on the extraction and analysis of appraisal groups such as 'very pretty', 'exceptionally good' which represents a set of attribute values in task-independent semantic taxonomies.

Li, S., Wang, Z., Zhou, G. and Lee, S.Y.M., (2011) have discussed the semi-supervised method to manage the lack of labeled data for sentiment classification by dividing the imbalance population into multiple sets of balanced populations and iterate it multiple times to improve the performance.

Wang, Y. and Wang, X.J., (2005) have discussed a variance mean based feature filtering technique that reduces the feature for the representative phrase of text classification in the Chinese language. This technique keeps only the best features to improve the overall performance and lessen the computing time.

Authors have used numerical matrix representation of movie review such that each column represents the extracted features, to solve this purpose (Tripathy, A., Agrawal, A., and Rath, S.K., 2015) applied count vectorizer and TF-IDF techniques and further apply Naïve Bayes and Support Vector Machine and resultant matrix provides the predicted positive and negative reviews.

Code-switching is a very popular phenomenon in the urban population. In India, millions of people use Hindi and English as popular form, some uses a pure form of these and many use a mixture of these languages, which is called Hinglish (Jhamtani, H., Bhogi, S.K., and Ray Choudhury, V., 2014) and proposed a model in taking care of the inconsistency in spelling while using Devnagiri for Roman.

Boiy, E. and Moens, M.F., (2009) determine sentiments for a specific entity in Web sentences (written in English, Dutch and French) mostly are loud and improperly framed. They discussed supervised machine learning techniques to lessen the quantity of preparing models by methods of dynamic learning, then test and evaluate the portability of the learned classification models across domains and languages.

Huang, F. and Yates, A., (2014) presented a model for LCS where firstly focuses on word alignment and pattern extraction to guide the EM training procedure for English and Chinese language. Secondly, they translate English words into Chinese and prepare a model.

Chandu, K., Loginova, E., Gupta, V., van Genabith, J., Neuman, G., Chinnakotla, M., Nyberg, E. and Black, A.W., (2018) discussed a model for code-mixing in Hinglish, Tamilish etc languages, which is used for crowdsourcing and prepared a question-answer set to look at the extent of lexical biasing towards CM.

Papadopoulos, F., Kitsak, M., Serrano, M.Á., Boguná, M. and Krioukov, D., (2012) discussed a method for making clusters based on deep learning algorithm by adding nodes based on popularity and similarity, where similarity is given higher priority in comparison to popularity.

The idea of this review is to present an impending of techniques majorly used by the researchers for sentiment analysis and a shift of user's tendency to express their views in multilingual rather than restricted to use a single language for expressions. Researchers have worked on sentiment analysis of unilingual posts, reviews, etc as well as doing work on the multilingual platform too. Due to constraints, (discussed in the below section) on the code-switching between Devanagri and English has restricted themselves to Hindi and English codes and didn't extend their work to Hinglish.

3. DESCRIPTION OF BROAD AREA

As the usage of the internet is increasing for expressing views among all categories of users and seeing the popularity of WWW among the multi-lingual community, WWW is also supporting various languages and various organizations are developing keyboards that make user free to type in their native languages as well as in English. Using the keyboards and various platforms, one can write one's views in a number of languages for expressing themselves, like Hindi, English, Spanish, French, Chinese, Japanese (Boiy, E. and Moens, M.F., 2009) etc. With the advancement, the usage of social networking is occurring in multiple languages. Mixing of multiple languages in a single sentence is very common among bi-lingual and multi-lingual communities. The mixing of multiple languages in a single sentence is called the Code-mixing phenomenon (Chandu, K. et al 2018). Code-

Mixing (CM) can be done by embedding phrases, words, and morphemes of one language into an expression of another language, which is commonly observed in multilingual communities (Myers-Scotton, C. 1997; Poplack,S., 1980; Muysken, P., Díaz, C.P. and Muysken, P.C., 2000; Chandu, K. et al 2018). This sometimes makes a feeling of lack of fluency in the segment of either language. Sometimes one can express his views by frequently switching from one language to another, to make his view, blog, speech, etc more impressive and impactful. For example, Indians may use Hindi words while speaking English (Huang, F. and Yates, A., 2014). For example, advertisements like "*Yahihai right choice baby"* use Linguistic Code-Switching to express and make an impact on a vast range of audiences. To find the opinion of public for a particular post, whether it is in favor or not, one needs to translate these languages into another and then need to check using uni-gram, bi-gram and n-gram techniques (Nene, S., 2017) for context analysis (according to their problems) whether this text is having positive expression or negative one.

3.1 Research Challenges

While translating from Devanagri (Hindi) to Roman (English) and vice versa, there can be a lot of challenges-

i) **Ambiguous words-** The words can be wrongly interpretive in translating from one language to another; because of ambiguous words usage like "bus" has a different meaning in English as well as in Hindi (Sinha, R.M.K. and Thakur, A., 2005).

ii) **Inconsistent spelling usage-** Although there are standards for transliterations are available, but people can use inconsistent spelling for the same word like the most common Hindi word "मैं" should be written as '*mai'* but some may write it as '*mei'* or '*mein'*. (Jhamtani, H. et al., 2014).

iii) **Some other issues are** developing a bag of words to provide a proper path of relationships among the words.

Due to these challenges many researchers (Jhamtani, H. et al., 2014; Chandu, K. et al., 2018; Sinha, R.M.K. and Thakur, A., 2005; Huang, F. and Yates, A., 2014) has restricted themselves to Hindi and English codes and didn't extend their work to Hinglish.

3.2 Problem Formulation

This work will concentrate on analyzing sentiments for multilingual languages Hindi, English, and Hinglish languages on social network data. The main focus will be lying on understanding the challenges and issues while working on Hindi, English, and Hinglish language and an approach will follow while performing sentiment classification for the user-generated content. With very few tools and annotated corpora available, it's a challenging task to perform sentiment analysis Hindi, English, and Hinglish languages altogether. The work will comprise of resource generation which involves the building of annotated datasets and subjective lexicon. An algorithm will be proposed to create Subjective Lexicon for the target languages using a pre-annotated seed list and wordnet of those languages. The generated result based on the target languages will be used to analyze the sentiment for the target event.

4. RESEARCH OBJECTIVES

The followings are the objectives of this study-

i) To study and understand the techniques of classification and clustering for classifying and clustering the sentiments of the Social Network.

ii) To build a corpus and subjective lexicon for Hindi and English from social Network data.

iii) To apply an algorithm to convert Hindi and Hinglish sentences in English Sentence.

iv) To implement a technique to build a subjective lexicon given a pre-annotated seed list for Hindi and English Language, its word set representing the network of words.

v) To design an algorithm for English Sentiment analysis for the targeted event.

vi) To analyze and evaluate the results based on the Hindi, English, and Hinglish Sentiment analysis for the targeted event and provides decisions that will be helpful to decide future steps.

5. RESEARCH METHODOLOGY

With the advent of WWW to support multilingual communication, the users of social media are restricted to communicate in English only. Users are using their mother tongues and

sometimes code-switching in a single utterance. Hindi is a sacred language and mother tongue of India also. In India, users use to communicate in Hindi, English and more often writing Hindi words in English or English words in Hindi which is called Hinglish. The proposed model will take into account these Hinglish words and develop a dataset. The developed annotated dataset is used for analyzing the sentiments from the web text; these sentiments, in turn, generate results and provide recommendations accordingly.

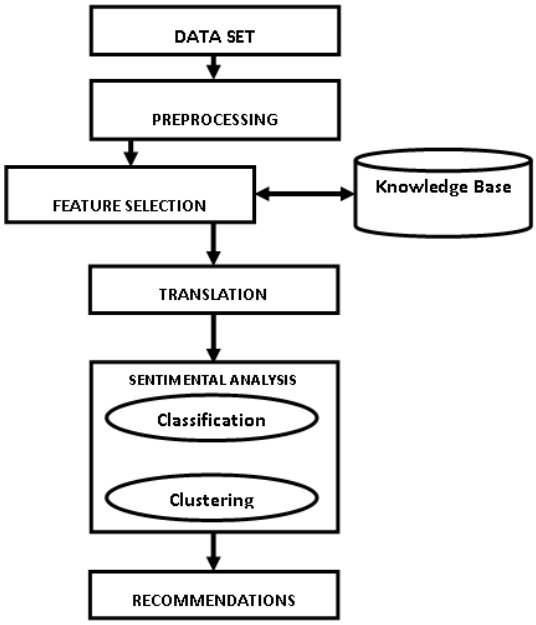

Figure 1: Methodology for Sentiment Analysis in Multilingual Social Network Data

i) **Data Set-** For performing sentiment analysis of the multilingual languages, a sufficient annotated dataset is not available. In this research, the following data sets will be the targeted one -

   a) Hinglish sentences

   b) English words frequency list

   c) Hindi words frequency list

ii) **Word level Language Identification-** To Identify similar words used in Hindi and English language for the same meaning, Frequency of occurrence ambiguous words in Hindi and English such as 'bus' has different meaning in both languages and based on n-gram e.g. beautiful, very beautiful, slightly beautiful or not beautiful all convey different meaning to the users; describe the most popular combinations of words.

iii) **Feature Selection** – To define the features based on which, the texts will be tokenized and segmented into sentences because the text is not always syntactically well-formed. For example: in Twitter user generally uses '@' symbol before using some name, use URLs to represent some information, etc. The target research identifies sentiments towards an entity across various sentences. Initially, sentiments will be uni-directional representing positive and negative attitudes.

iv) **Translation**- The translation should be performing to convert Hinglish words into Hindi or English, tag the words for the language it is using.

v) **Machine Learning Techniques-** Many researchers are working on sentiment analysis and are using Support Vector Machine (Tripathy, A., Agrawal, A. and Rath, S.K., 2015) proposed TF-IDF in combination to SVM to find a highly accurate data, whereas (Turney, P.D., 2002; Boiy, E. and Moens, M.F., 2009) are using entropy-based classifiers to classify the text and other are using Naïve Bayes Technique (Tripathy, A., Agrawal, A. and Rath, S.K., 2015; Pang, B., Lee, L. and Vaithyanathan, S., 2002; Boiy, E. and Moens, M.F., 2009). However, these techniques are simple and provide a good result.

vi) **Recommendations-** Recommendation system is a sub-class of information filtering such that the system can predict the rating or preferences to the dataset given by the user and can provide feedback to increase the performance of the datasets having a negative rating.

6. EXPECTED OUTCOME OF THE RESEARCH

i) Development of a corpus and subjective lexicon for pre-annotated seed list for Hindi and English Language.

ii) Development of the system for Sentiment Analysis based on Multilingual Social Network Data which help to increase the performance of target event.

REFERENCES

Bauer, F., Pereverzev, S. and Rosasco, L., 2007. On Regularization Algorithms In Learning Theory. *Journal Of Complexity*, *23*(1), pp.52-72.

Beringer, J. and Hüllermeier, E., 2007. An Efficient Algorithm For Instance-Based Learning On Data Streams. In *Industrial Conference On Data Mining.* pp. 34-48.

Beyer, M.A., and Laney, D., 2012. The importance of 'big data': a definition. *Stamford, CT: Gartner*, pp.2014-2018.

Bhaskaran, K., Gasparrini, A., Hajat, S., Smeeth, L. and Armstrong, B., 2013. Time Series Regression Studies In Environmental Epidemiology. *International Journal Of Epidemiology*, *42*(4), pp.1187-1195.

Boiy, E. and Moens, M.F., 2009. A Machine Learning Approach To Sentiment Analysis In Multilingual Web Texts. *Information Retrieval*, *12*(5), pp.526-558.

Bringmann, B., Berlingerio, M., Bonchi, F. and Gionis, A., 2010. Learning And Predicting The Evolution Of Social Networks. *IEEE Intelligent Systems*, *25*(4), pp.26-35.

Chandu, K., Loginova, E., Gupta, V., van Genabith, J., Neuman, G., Chinnakotla, M., Nyberg, E. and Black, A.W., 2018. Code-Mixed Question Answering Challenge: Crowd-Sourcing Data And Techniques. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching,* pp. 29-38.

Cord, M. and Cunningham, P. eds., 2008. Machine Learning Techniques For Multimedia: Case Studies On Organization And Retrieval. Springer Science & Business Media, pp.3-20.

Cortés, G., Benitez, M.C., García, L., Álvarez, I. and Ibanez, J.M., 2015. A Comparative Study Of Dimensionality Reduction Algorithms Applied To Volcano-Seismic Signals. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *9*(1), pp.253-263.

Dave, K., Lawrence, S. and Pennock, D.M., 2003. Mining The Peanut Gallery: Opinion Extraction And Semantic Classification Of Product Reviews. In ACM *Proceedings of the 12th international conference on World Wide Web,* pp. 519-528.

Diebold, F.X., Cheng, X., Diebold, S., Foster, D., Halperin, M., Lohr, S., Mashey, J., Nickolas, T., Pai, M., Pospiech, M. and Schorfheide, F., 2012. A Personal Perspective on the Origin (s) and Development of "Big Data": The Phenomenon, the Term, and the Discipline, pp.03-13.

Doan, T. and Kalita, J., 2015. Selecting Machine Learning Algorithms Using Regression Models. In *2015 IEEE International Conference on Data Mining Workshop (ICDMW),* pp. 1498-1505.

Domingos, P. and Pazzani, M., 1997. On The Optimality Of The Simple Bayesian Classifier Under Zero-One Loss. *Machine Learning*, *29*(2-3), pp.103-130.

Elgendy, N. and Elragal, A., 2014. Big Data Analytics: A Literature Review Paper. In *Industrial Conference on Data Mining,* Springer, Cham, pp. 214-227.

Fan, J., Han, F., and Liu, H., 2014. Challenges Of Big Data Analysis. *National Science Review*, *1*(2), pp.293-314.

Friedman, J.H. and Roosen, C.B., 1995. An Introduction To Multivariate Adaptive Regression Splines,pp.-197-217.

Gandomi, A. and Haider, M., 2015. Beyond The Hype: Big Data Concepts, Methods, And Analytics. *International journal of information management*, *35*(2), pp.137-144.

Greene, W., 2009. Discrete Choice Modeling. In *Palgrave Handbook of econometrics,* Palgrave Macmillan, London, pp. 473-556.

Huang, F. and Yates, A., 2014. Improving Word Alignment Using Linguistic Code-Switching Data. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics,* pp. 1-9.

Jha, A., Dave, M. and Madan, S., 2016. A Review On The Study And Analysis Of Big Data Using Data Mining Techniques. *Int J Latest Trends Eng Technol (IJLTET)*, *6*(3), pp.94-102.

Jhamtani, H., Bhogi, S.K., and Raychoudhury, V., 2014. Word-Level Language Identification In Bi-Lingual Code-Switched Texts. In *Proceedings of the 28th Pacific Asia Conference on Language, Information and Computing,* pp.348-357.

Lesser, V., Horling, B., Klassner, F., Raja, A., Wagner, T. and Zhang, S.X., 2000. BIG: An Agent For Resource-Bounded Information Gathering And Decision Making. *Artificial Intelligence*, *118*(1-2), pp.197-244.

Lewis, R.J., 2000. An Introduction To Classification And Regression Tree (CART) Analysis. In *Proceedings of Annual Meeting of the Society for Academic Emergency Medicine in San Francisco, California*, 14, pp.1-14.

Li, S., Wang, Z., Zhou, G. and Lee, S.Y.M., 2011. Semi-supervised learning for imbalanced sentiment classification. In *Twenty-Second International Joint Conference on Artificial Intelligence,* pp.-1826-1831.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C. and Byers, A.H., 2011. Big data: The Next Frontier For Innovation, Competition, And Productivity,pp.1-13.

Mujeeb, S. and Naidu, L.K., 2015. A Relative Study On Big Data Applications And Techniques. *International Journal of Engineering and Innovative Technology*, *4*(10), pp.133-138.

Musumeci, F., Rottondi, C., Nag, A., Macaluso, I., Zibar, D., Ruffini, M. and Tornatore, M., 2018. An Overview Of Application Of Machine Learning Techniques In Optical Networks. *IEEE Communications Surveys & Tutorials*, *21*(2), pp.1383-1408.

Muysken, P., Díaz, C.P. and Muysken, P.C., 2000. *Bilingual Speech: A typology of code-mixing* Cambridge University Press, pp. 250-270.

Myers-Scotton, C., 1997. *Duelling languages: Grammatical structure in code-switching*. Oxford University Press.

Nene, S., 2017. Deep Learning for Natural Language Processing. *International Research Journal of Engineering and Technology (IRJET), 4* (11), pp.930-933.

Pang, B. and Lee, L., 2004. A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based On Minimum Cuts. In *Proceedings Of The 42nd Annual Meeting On Association For Computational Linguistics,* pp. 271-278.

Pang, B., Lee, L. and Vaithyanathan, S., 2002. Thumbs Up?: Sentiment Classification Using Machine Learning Techniques. In *Proceedings Of The ACL-02 Conference On Empirical Methods In Natural Language Processing,* 10*,* pp. 79-86.

Papadopoulos, F., Kitsak, M., Serrano, M.Á., Boguná, M. and Krioukov, D., 2012. Popularity Versus Similarity In Growing Networks. *Nature*, *489*(7417), pp.537.

Phaneendra, S.V., and Reddy, E.M., 2013. Big Data-solutions for RDBMS problems-A survey. In *12th IEEE/IFIP Network Operations & Management Symposium (NOMS 2010),*2(9), pp. 3686-3691 .

Poplack, S., 1980. Sometimes I'll Start A Sentence In Spanish Y Termino En Espanol: Toward A Typology Of Code-Switching1. *Linguistics*, *18*(7-8), pp.581-618.

Sagiroglu, S. and Sinanc, D., 2013. Big Data: A Review. In *2013* IEEE *International Conference on Collaboration Technologies and Systems (CTS,),* pp. 42-47.

Singh, Y., Bhatia, P.K. and Sangwan, O., 2007. A Review Of Studies On Machine Learning Techniques. *International Journal of Computer Science and Security*, *1*(1), pp.70-84.

Sinha, R.M.K. and Thakur, A., 2005. Machine Translation Of Bi-Lingual Hindi-English (Hinglish) text. *10th Machine Translation summit (MT Summit X), Phuket, Thailand*, pp.149-156.

Tamhane, D.S. and Sayyad, S.N., 2015. Big Data Analysis Using Hace Theorem. *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, *4*, pp.2278-1323.

Tripathy, A., Agrawal, A., and Rath, S.K., 2015. Classification of Sentimental Reviews Using Machine Learning Techniques. *Procedia Computer Science*, *57*, pp.821-829.

Turney, P.D., 2002. Thumbs Up Or Thumbs Down?: Semantic Orientation Applied To Unsupervised Classification Of Reviews. In *Proceedings of the 40th Annual Meeting On Association For Computational Linguistics,* pp. 417-424.

Wang, Y. and Wang, X.J., 2005, August. A New Approach to Feature Selection In Text Classification. In *2005* IEEE *International Conference On Machine Learning And Cybernetics,* 6, pp. 3814-3819.

Whitelaw, C., Garg, N. and Argamon, S., 2005. Using Appraisal Groups For Sentiment Analysis. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management,* pp. 625-631.

Zou, K.H., Tuncali, K. and Silverman, S.G., 2003. Correlation and Simple Linear Regression. *Radiology*, *227*(3), pp.617-628.

Neha_Garg_Short_Synopsis_NG_05.04.19.

ORIGINALITY REPORT

9%	6%	8%	2%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

Excluding matches less than 14 words

06/04/19

Neha Garg 06/04/19

6/04/2019

PRIMARY SOURCES

1	Tripathy, Abinash, Ankit Agrawal, and Santanu Kumar Rath. "Classification of Sentimental Reviews Using Machine Learning Techniques", Procedia Computer Science, 2015. Publication	2%
2	machinelearningmastery.com Internet Source	2%
3	documento.site Internet Source	1%
4	Submitted to Savitribai Phule Pune University Student Paper	1%
5	ethesis.nitrkl.ac.in Internet Source	1%
6	Abinash Tripathy, Abhishek Anand, Santanu Kumar Rath. "Document-level sentiment classification using hybrid machine learning approach", Knowledge and Information Systems, 2017 Publication	1%

Dr. Vikram Kr. Sharma
University Librarian
Dr. O.P. Bhalla Central Library
Manav Rachna International Institute of Research & Studies
Faridabad

7	Submitted to General Sir John Kotelawala Defence University Student Paper	1 %
8	"Information Systems Design and Intelligent Applications", Springer Nature, 2018 Publication	<1 %

Exclude quotes	On	Exclude matches	< 14 words
Exclude bibliography	On		